DOCUMENT RESUME

ED 480 036 CG 032 609

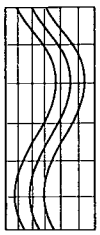
AUTHOR	Fremer, John; Wall, Janet
TITLE	Why Use Tests and Assessments?
PUB DATE	2003-08-00
NOTE	19p.; In: Measuring Up: Assessment Issues for Teachers, Counselors, and Administrators; see CG 032 608.
PUB TYPE	Information Analyses (070)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Educational Assessment; Educational Environment; *Educational Testing; Elementary Secondary Education; *Evaluation Methods; Instructional Effectiveness; Standardized Tests; Testing Problems; Theory Practice Relationship

ABSTRACT

This chapter outlines the purpose of testing and assessment, focusing on uses, and highlights some of the limitations of all forms of testing. The concept of testing is one of the major contributions of the field of psychology to society. Carefully developed tests, when used wisely, provide valuable information for decision makers in educational, employment, and clinical settings. It is because of their often-demonstrated utility that tests and other standardized assessments are so widely used in educational settings. In order to gain the potential benefits that tests offer, it is essential to be aware of their strengths and their limitations. In this chapter, the following key aspects of high-quality testing are reviewed: what is a test or assessment; what are the major uses of tests; what are the key benefits of systematic, high-quality testing; what are the frequent criticisms of testing; and how can we promote high-quality testing. (Contains 14 references.) (GCP)

Why Use Tests and Assessments?

By
John Fremer
Janet Wall



Chapter 1
Why Use Tests and Assessments?
Questions and Answers

John Fremer & Janet E. Wall

The terms *assessment, measurement,* and *testing* will be used heavily in this book. Although the terms are often used interchangeably, there are some distinctions between them. *Testing,* generally considered to be the most narrow or specific of the terms, tends to refer to a set of questions that has been compiled to measure a specific concept such as achievement or aptitude. *Assessment* is broader in scope; it encompasses testing, but can also include measurement via observations, interviews, checklists, and other data gathering instruments. The term assessment is used more often in the clinical setting or for determining preferences, interests, and personality types. The term *measurement* generally refers to the attempt at quantifying the results of tests and assessments. This chapter will outline the purpose of testing and assessment, focus on uses, and highlight some of the limitations of all forms of testing.

The concept of testing is one of the major contributions of the field of psychology to society. Carefully developed tests, when used wisely, provide valuable information for decision makers in educational, employment, and clinical settings. It is because of their often-demonstrated utility that tests and other standardized assessments are so widely used in educational settings. In order to gain the potential benefits that tests offer, it is essential to be aware of their strengths and their limitations. In this chapter, we review these key aspects of high-quality testing:

- What is a test or assessment?
- What are the major uses of tests?
- What are the key benefits of systematic, high-quality testing?
- What are the frequent criticisms of testing?
- How can we promote high-quality testing?

What Is a Standardized Test or Assessment?

During the medieval period in Europe, skilled craftsmen who were members of a guild carried with them symbols of their trade. We do not

have many examples of that practice now, but the stethoscope around a doctor's neck, the chalk in the hands of a teacher, or the tool belt of a carpenter or telephone line worker all bring to mind that person's line of work. What might a tester carry to signal his or her professional role? It could be a copy of the Iowa Test of Basic Skills or the Florida Comprehensive Assessment (FCAT). Perhaps the Myers-Briggs Type Indicator (MBTI) or the Minnesota Multiphasic Personality Inventory (MMPI). What about a driver's test or a military entrance exam? Yet other options could be an SAT-I, advanced placement test, or a copy of an ACT Assessment.

Basically, testing is a special way of collecting information used to help make decisions about individuals, programs, or institutions. Tests and assessments are generally made up of items or questions that elicit responses from an individual. It is important to note that merely administering some set of questions or performance tasks is only one part of the testing process. If the tests are never scored and the results never used, we have done only part of what is needed. Yet there are instances ranging from the individual classroom level to nationwide assessment where tests are given and little use is ever made of the information. In order for actual measurement to take place as part of testing, one or more of the following steps must take place:

- An individual or group must receive a score along with some guide to interpreting that score.
- The individual or group must be ranked against others who have been tested.
- The individual or group must be classified into some meaningful category; for example, "gifted," "shows some evidence of obsessive behavior," "merits a personal interview," or "needs further evaluation."
- The performance of the individual or group must be compared against some explicit standard.

Most instances of testing very clearly meet one or more of these criteria: An individual who takes a required test receives a score on a well-defined scale and also receives a good deal of comparison information and an interpretive guide. In other instances the situation is not so straightforward. For example, a teacher asks the class to answer a set of questions and to send in electronic or paper responses. The teacher reads all the responses and makes a judgment as to how well the group as a whole has learned the material covered by the questions. Has measurement taken place? Yes, for the class as a unit, but no for the individuals, if the teacher has not classified their responses in any

way. In real life, of course, the teacher may recall the specific responses of some students and either confirm or change his or her perception of their level of understanding. For that subset of students the testing process has actually led to measurement. The issue, "What is measurement?" is reviewed by Jones (1971), who notes that although "unanimity concerning the meaning of measurement may appear unlikely . . . each measurement is purposive . . . and the purpose is always . . . to acquire information" (p. 335).

What Are the Major Uses of Tests?

We have maintained that the basic purpose of tests is to provide information for decision makers. In the last section we made the case that the process must also include assigning a score, rank, or classification of some type. We now want to describe five major uses of test results, as follows:

- selection or placement
- diagnosis
- accountability evaluations
- judging progress and following trends
- self-discovery

Selection or Placement

The use of tests to help select individuals for admissions to an institution or special program is so widespread that it is perhaps best described as a standard feature of U.S. society. Entrance examinations are used as early as entrance into kindergarten and with increased frequency as the student moves up the grades and into college and a profession. Usually test information is combined with grades to make decisions; tests are also frequently used at the college level to grant exemption from or credit for college courses taken while a student is still attending high school (Willingham, Lewis, Morgan, & Ramist, 1990).

When using a test to help make selection or placement decisions, it is essential that the decisions made be of higher quality when the tests are used than when they are not. If students are being accepted for admission to a college, for example, the group that is admitted should perform better than the group that would have been chosen without the use of tests.

How might we determine whether tests had improved our process for selecting college students? We could look at overall grade point

average, grades in specific courses, record of successful completion of the freshman year, or persistence to graduation of the students who were admitted. Each of these criteria has been employed to evaluate the value of college admissions tests. Most often, though, it is freshman grade point average (FGPA) that is employed in studies of the value of the SAT I and SAT II and of the ACT Assessment. FGPA is routinely determined by virtually all colleges, so it is an easy bit of criterion information to obtain. The results from many thousands of studies of the value of college admissions have yielded consistent results. For most colleges, high school grades are the best predictor of college grades (Donlon, 1984). For many colleges, though, admissions test scores are the best single predictors. The most common practice is to use both test scores and high school grades. Increasingly, colleges are looking at all available information about students, including recommendations, personal essays, past accomplishments, community service record, and other evidence of a student's potential for college achievement and for subsequent contributions to society. In a classic work on this topic, Willingham & Breland (1982) point out that although "some personal qualities are related to success, some have intrinsic merit in their own right and some are demonstrably related to important institutional objectives" (p. 3).

Whereas a great deal of study has been devoted to evaluating the strengths and limitations of college admissions tests, much less attention has been given to other uses for tests in educational selection and placement settings. It is very common for one or more tests to be used to select students for gifted and talented or special education programs. Ideally the managers of such programs would first define the student characteristics that each program is designed to nurture and develop. Then they would develop selection procedures to choose the most appropriate students for the program. Some combination of prior academic work, teacher recommendations, and test results will typically be most effective in the selection process. Whatever approach is used, the results should be carefully evaluated to make sure that all the information used is having the desired contribution to picking the group of students for whom the program will be most effective.

Diagnosis

Tests are also used extensively to evaluate students' special needs. Test results help educators, counselors, and other professionals plan individualized education programs for students or point out specific misconceptions or problem areas that are hindering progress. Often

tests help determine the need for counseling services, especially when students are experiencing high personal stress or engaging in substance abuse or other harmful and dangerous behaviors. The home and workplace are other contexts where physical and psychological problems occur for which tests are often part of the solution.

Some of the tests used in diagnostic settings in education measure basic academic skills and knowledge. Has a child mastered basic linguistic and mathematical content? If not, what are the child's areas of strength and weakness? Often a classroom teacher will ask for special diagnostic testing for a child who is not keeping up with other students or not responding to the teaching methods being employed. The goal of diagnostic educational testing of this type is to add information that can be used to plan the child's future educational program. The closer the test content is to the skills that are the goals for instruction, the more useful the results will be to those doing such planning. Another consideration in evaluating the results of diagnostic educational tests is the extent to which parents and other family members can readily understand the test results.

In addition to skills-oriented diagnostic educational tests, trained professionals frequently use a number of other tests, surveys, and inventories in educational and employment settings. Successful performance in school, work, or other settings is not merely a question of having the necessary skills. Grief, anxiety, anger, and other debilitating states of mind can have strong effects on children and adults of all ages and in many different life situations. Tests and related tools can help focus attention on the nature and extent of the difficulty that is interfering with the individual's ability to perform effectively.

The interpretation of nonacademic tests requires specialized training in areas such as counseling, school psychology, or clinical psychology. The trained professional takes into account many factors of a person's situation in order to evaluate test results in a proper context and to make useful recommendations (Bracken, 1991).

Accountability Evaluations

Some testing in education is carried out for the express purpose of holding students, educators, and schools accountable for their performance. Such testing programs set explicit standards and require adherence to these standards, often with some form of reward for those who achieve them and sanctions for those who do not. For students, a requirement to attend summer school is one possible consequence of failing to meet a grade promotion standard. Retention in the current

grade or failure to graduate from high school are other possible results of failing to meet the standard.

For teachers, a possible positive outcome is a cash bonus based on high student performance. A possible negative consequence for a school or school system is loss of autonomy, a takeover by a higher administrative unit. In each instance, the goal is to ensure adequate and predefined levels of student performance, as measured by a particular set of tests designed or selected for that purpose. A basic assumption of accountability testing programs is that the identification of performance targets, with testing and consequences for results, will lead to more focused instruction and higher performance.

A major milestone in the growth of accountability testing programs occurred in 2002, when Congress passed the No Child Left Behind (NCLB) act. This legislation requires regular accountability testing as a prerequisite for receiving federal funds. The legislation puts pressure on states to identify poorly performing schools and seek remedies to the situation. The legislation requires testing of higher-order skills and the alignment of assessment to state standards.

Advocates of this dramatic expansion characterize it as an essential step in obtaining systematic information on the effectiveness of educational programs for all students. Critics of the new federal requirement for testing object to what they characterize as a one-size-fits-all overemphasis on testing of basics to the detriment of other important aspects of school programs.

All a testing program can do is collect information and summarize it for those who can use it. In order to evaluate the merits of arguments for and against educational accountability testing programs of various types, one needs to ask incisive questions about the purposes, procedures, and outcomes of these programs: What information is being collected and at what stage in a child's education or a school's program? Is there a good match between what is being taught and what is being tested? Is understandable information being provided to appropriate people, including students and parents, in a timely fashion?

When an accountability program is soundly designed and executed, it can be a valuable component of the educational process. In many ways such a regular checkup on the effectiveness of an educational program is as valuable as the health exams that we seek periodically and the independent financial audits that companies should receive regularly. In every instance we want to see the proper tests employed, but the real payoff comes from skilled interpretation and proper follow-up based on the results.

Judging Progress and Following Trends

In addition to providing information about the group that is currently being tested, an ongoing testing program permits comparisons over time. An important example in American education is the National Assessment of Educational Progress. The word *progress* in the official name of this program is no accident. The intention of American educators is to ensure that students develop their skills as they proceed through school and that improvement from year to year occurs in the overall performance of students and schools. Testing provides a way of describing the current status of education and of tracking trends over time.

One of the significant developments in testing during the later years of the twentieth century was the increase in public attention to the results of international studies of education. These cross-national projects have tended to focus on basic subjects such as reading or language arts and mathematics. There have been significant controversies within the United States as to what the results tell us. It seems quite clear, though, that our students are a far step from being "first in the world in mathematics and science," a national educational goal for the year 2000. Part of the difficulty in evaluating cross-national comparisons is that the vast majority of our elementary and secondary students remain in school until at least age 16. In many countries, most students end their formal education before that time. Very different results are found if we compare the mathematics performance of all our high school students, versus only those in advanced placement mathematics courses, with the performance of students from other countries. The outcomes could be called either very worrisome or exemplary, depending on the U.S. comparison group.

As with any type of trend data, the real benefits accrue as data are collected over several years. Longitudinal data allow us to answer questions such as, Are we doing a better job now than we were the last time we checked? How well are we meeting the needs of the many subgroups that make up our society? Part of the substantial value of well-crafted standardized tests is that they can help us answer questions of this type (Ekstrom & Smith, 2002; Willingham & Cole, 1997; Zwick, 2002). Progress on the individual student level is critical as well. School educators and parents expect to see students making yearly progress due to the provision of an excellent education program offered at the schools.

Self-Discovery

There are many important aspects of people beyond the domains of skills and competencies. It should not be surprising, then, that there are a wide variety of tests in areas such as attitudes, motivation, personality, and other psychological characteristics. With the assistance of a trained testing professional—such as a counselor, psychologist, social worker, or member of some other relevant helping profession—individuals can gain information to help them make more informed career and life decisions as well as deal with troublesome life circumstances. We need only scan the tables of contents of general interest magazines or browse the Internet to find tests that purport to help us find a mate, choose an ideal line of work, or achieve a deeper understanding of "who we really are." Since these tests may not meet high professional standards, they should not be taken too seriously. They may indeed help us reflect on aspects of who we are and how we view the world. If we are facing critical decisions, though, or dealing with some problem that is interfering with our relationships, work, or ability to lead the life we want to live, a test alone is not likely to meet our needs. A skilled professional can combine test results with other information about us and our situation, and work with us to help us improve our quality of life.

When evaluating the results of tests that are designed to help someone gain self-understanding, it is important to take into account the issues of honesty and consistency. Turning first to honesty, keep in mind that the test outcomes will depend in good part on how accurately the person describes himself or herself. If one always chooses the answer that describes an ideal person's choice, rather than one's own, there is no reason to expect an accurate result and interpretation. If someone responds as though he or she never had a selfish thought in their life, and would always choose to visit a sick friend over meeting a favorite athlete, singer, or movie star, don't be surprised if the resulting profile seems a lot nobler than the person really is.

It is also important to consider whether the results from testing are consistent with other available information. This is especially important when making important life decisions. Weigh the results of testing along with the many other pieces of information already available. Be open to new insights, surely, but wonder and seek advice about any guidance that seems contradictory to everything else known. For a true "city" person with little interest in the natural world, perhaps a job as a forest ranger or a gardener may not be as ideal as for someone who finds cities a noisy irritant to life.

What Are the Key Benefits of Systematic, High-Quality Testing?

Given that decision makers are seeking information to help them make decisions of the type we reviewed in the last section, what are the benefits of using the results of high-quality testing? We will discuss each of these valuable qualities:

- objective results
- cost-effectiveness
- technical quality and standards
- fairness
- evolutionary improvement

Objective Results

High-quality testing produces objective results, that is to say results that are consistent from occasion to occasion. This outcome is very clear when you use multiple-choice or other machine scored tests. You can score a test twice and get the same result. High levels of objectivity can also be obtained with assessment exercises that require professional scoring. It is essential to use exercises that well-trained and monitored scorers can grade with the necessary level of consistency, but this is a challenging, albeit attainable, goal. Whatever the type of test to be developed and used, issues such as the following need to be carefully addressed:

- What is the purpose of the test, and how will the results be used?
- What are the characteristics of the people who will take the test?
- What areas of content and skill will be measured?
- What scoring rules will be followed and how will accuracy be checked?

Cost-Effectiveness

High-quality tests are among the most cost-effective means available to obtain high-quality information. To obtain consistent measures of a student's or worker's performance outside of a specially designated testing situation requires several observations and two or more judges or observers. Such observations in real-life situations frequently involve 10 to 100 times the cost to achieve the same level of exactitude as a standardized test.

Technical Quality and Standards

The growth of the standardized movement in the United States has been accompanied by the development and refinement of professional standards for testing. Major testing companies and professional associations whose members make frequent use of tests have endorsed the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) as well as the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2002). Users of test results need to ascertain whether the tests being reported have been prepared in accordance with these standards. If so, it is possible to place considerable confidence in the technical quality of the results.

Fairness

Test makers strive to develop tests that primarily reflect the skill, knowledge, or other characteristics the test is intended to measure. If individuals or groups differ with respect to what is being measured, then test results should reveal those differences. Unfairness occurs, for example, when factors extraneous to the skill being evaluated have a significant influence on test scores. A test of mathematical skill that uses complex language or sets problems in contexts unfamiliar to test takers would not be a fair measure of this particular skill. Helms (1997) looks at the interaction of race, culture, and social class in cognitive ability testing and concludes that advisories should be included when reporting test scores of test takers with experiences and backgrounds different from advantaged students.

It is important to note that a fair test result, in the sense that it accurately portrays the competence of an individual or group, may well be perceived as unfair. For example, a parent who very much wants to see his or her child admitted to a highly selective program, school, or college will tend to reject any indicator that does not contribute to this objective (Zwick 2002).

Evolutionary Improvement

One of the great strengths of standardized testing as it is typically carried out in the United States by professional test developers is that the basic approaches employed support evolutionary improvement in the quality of the tests. Whereas we may say that we are going to review our work at the close of any important project in our work or home life, all too often we move on to the next task without systematically re-examining the extent to which we attained our goals. Life seems to

bring us one string of demands after another and finding the time to review carefully the job just completed loses out to the need to go forward to the next challenge.

In standardized testing, reviewing how we did with this year's test is an essential part of the craft and science of test making. No credible testing professional fails to carry out thorough analyses of a test once it is administered, scored, and reported or interpreted. Indeed the professional standards governing the work of test developers and test users require this type of systematic review. Test developers aim for a test at a certain level of difficulty and one that has scores meeting a particular level of *reliability,* the test makers' term for consistency of results. Text makers explore questions appropriate to the test and how it is used through analyses of data collected at the time of testing and, where necessary, through additional data collection steps. For example, did we achieve our difficulty-level and reliability goals? Moreover, if the test is being used to select people for academic programs or for hiring or promotion, what is the evidence that the test provides information to support such decisions?

This practice of collecting and analyzing data over the years to evaluate the effectiveness of tests and test use is called *validation,* and it provides a regular means for improving the quality of a test over the period of its use. It is this phenomenon that we refer to as evolutionary improvement, and it is a quite important feature of standardized tests when wisely planned and used for test and test program improvement.

What Are the Frequent Criticisms of Testing?

A clear pattern to test criticism emerges if one takes the time to read through the many years of discussions of testing in popular news magazines, leading newspapers, and other major publications. Each time there is a substantial increase in the use of standardized testing or a new application of the approach, a wave of criticism follows. If the application is well planned and implemented, the intensity and duration of the criticism may be lessened, but the critics will still insist on being heard. When a new testing application has been introduced with insufficient notice to those affected or in a manner that violates professional testing standards and good sense, there can be prolonged and very strong criticism that on occasion stalls or derails the proposed testing application. Haney, Madaus, & Lyons (1993) include a chapter, "Test Quality and the Fractured Marketplace for Testing," that provides one perspective on the influence of market factors on test quality.

In this section we are going to look at four main classes of test criticism. A large percentage of all the concerns that are voiced about testing can be classified into one of these four categories:

- bias versus fairness
- coachability
- appropriateness of use
- technical quality

Bias Versus Fairness

The issue of whether standardized tests are fair or promote fairness in decision making is one of the most persistent and, on occasion, hotly debated areas of criticism of testing (Leman, 1999; Willingham & Cole, 1997; Zwick, 2002). One of the reasons for the heat and emotion that frequently characterize discussions of this issue is that the debate has a personal component. Those involved may begin with academic and dispassionate statements about the results of a particular testing program, but they often end up addressing the impact of testing on their own lives or those of their children or other family members.

Communication on the issue of bias or lack of it in testing is also complicated by the fact that participants in the discussion typically bring different definitions of bias to the situation. As we noted in our earlier discussion of test fairness, testing professionals examine the possibility that a test may be biased by checking to see if the results from that test are consistent with other information about the group being tested. If we look at the types of individuals earning high or low scores, does this result make sense in light of other information about the competencies measured by the test? For example, we would expect students who excel in their math classes, have joined the math club, and work on math puzzles for recreation to score high on a math test. Similarly, we would expect those students who take the absolute minimum number of math courses and ignore or fail to deliver math homework not to perform well.

Coachability

Groups reflecting a variety of perspectives would likely view evidence that scores could be readily affected by short-term coaching as a troublesome feature of any test. Thoughtful observers would also worry if a test were much better at predicting the future success of some groups of students than of others. Our expectation and requirement for tests is that they will be effective for all citizens, not only for a subset. Another issue in evaluating coaching are the definitions of

coaching and short-term test preparation. In some areas, such as mathematics, there are effective short-term seminars that build both test-taking and other content skills. Finding that scores go up as the underlying skills being tested improve is a good feature of a test, not a source for concern.

In studies of the SAT I and the ACT Assessment, the tests for which perhaps the largest number of studies of the effectiveness of coaching have been carried out, only a small contribution to future test scores can be attributed to coaching courses (Messick, 1980). Whereas individuals trained in testing are inclined to be persuaded by these research results, the coaching companies report gains that are far greater than these average results. Often they present glowing testimonials to the effectiveness of coaching for an individual student, usually without much detail about the circumstances under which the reported gains were obtained. One of our speculations is that coaching companies look only at the gains made by those who earn higher scores, leaving out of their calculations those whose scores stay the same or decrease. So the "average gain" they report is actually the average gain of those who gained, not the average gain across all those who were taught.

As to advice for those who face standardized tests or are working with students or others who take them, by all means prepare carefully for any important test. Read the available material about the test, especially that produced by the test makers. Become very familiar with the types of questions you will encounter. Don't waste time figuring out what you need to do on the day of the test when others have done this task weeks or months ago. If everyone who comes to a test has done this type of preparation, the force of a criticism on the grounds of coachability is substantially undermined.

Appropriateness of Use

One of the criticisms of standardized testing that seems to us to be well supported in many instances deals with the use to which a test is put. Professional test developers are charged by their standards to be explicit about what the intended uses of their tests are. Agencies that choose to employ a test for a different purpose than that for which the test was developed have a responsibility to document the appropriateness of the test for the new use. For example, as we noted earlier in this chapter, a test designed to measure the mathematics competence of a native-speaking group may be completely inappropriate for judging the level of mathematical knowledge of many non-native speakers of English. They could have the skill that is intended to be

measured but be unable to show their competence because of their inability to understand the problems that were set for them.

Technical Quality

Criticisms about the technical quality of tests sometimes focus on individual test questions, asserting that they are inadequate for the purpose of testing. Particular items may be judged too easy or too difficult, or perhaps described as too low-level or too ambiguous. The critic may assert that the coverage of the test is too shallow or is unbalanced in some way, giving too much weight to one or more areas and slighting another topic or topics. The format may be dismissed as wrong for the test in question; this happens especially with multiple-choice questions, although sometimes with essay or other question types. Another type of criticism regards the number of questions and the stability of the resultant score.

Those involved in the selection and use of tests should take the concerns of critics seriously, either evaluating the criticisms for themselves or bringing other trained professionals into the process. In some instances we stand to learn more by listening to critics than we would gain by proceeding to enjoy the support of friends, who may give us the benefit of the doubt and fail to give needed criticism.

How Can We Promote High-Quality Testing?

One of the most important ways to promote high-quality testing is to become familiar with the types of testing that are going on. How are tests being used? Are the purposes clearly stated, and do the kinds of tests being employed seem consistent with those purposes? What issues are being raised by individuals who find fault with the testing? Do the criticisms seem warranted?

One of the many benefits of our Internet-linked world is that it is now quite easy to look up what is being said about any test or testing program. We urge looking at both sides of any testing issue. Just as the maker of a test is predisposed to see the virtues of the product that is produced, some critics reflexively reject virtually any standardized test, no matter how carefully it is crafted and how educationally or occupationally sound the use to which it is being put.

For many people being an effective evaluator and user of tests and test results will require learning more about tests and test making. This book is one way for individuals to expand their knowledge. Chapter

53 provides guidance on accessing many types of resources. The following list provides some additional sources of information about tests and test quality.

Sources of Information

One of the best sources of information about any test or testing program is the test developer. Descriptive material, registration bulletins, score reports, and other test-related documents are often available from the test publisher. Be sure to take advantage of any information that you can obtain directly from the publisher; it is usually accurate and up-to-date, and it is often free.

Even if you have no printed materials in your possession, a systematic search of the Internet will often be very productive. Look for test descriptions, sample questions, and media coverage of the tests. There are two major trade associations of test publishers and one professional association, all of whose websites are resources both for general information and for locating specific test publishers and other testing agencies.

Association of Test Publishers (ATP; www.testpublishers.org). This association has well more than 100 member companies, representing clinical, educational, employment, and licensing/certification areas. This association is particularly active in the area of computer-based testing.

Association of American Publishers (AAP; www.publishers.org). This association represents all or virtually all the publishers of textbooks, tests, and related material for U.S. schools and colleges. The AAP Test Committee plays an active role in monitoring legislation and regulations related to testing. It has produced several fine publications about testing.

Joint Committee on Testing Practices (JCTP; www.apa.org/science/jctpweb.html). JCTP is a collaboration among professional associations whose members make extensive use of tests and testing companies. The organization's goal is to work "together to advance in the public interest the quality of testing practices" (JCTP home page, accessed 1/23/03). The JCTP has produced a number of helpful publications and other materials covering areas such as testing standards, test-purchaser qualifications, teaching about testing, the rights and responsibilities of test takers, and the testing of individuals with disabilities.

Collecting Specific Documents

If you are in a position to serve as a resource for others in the area of testing, you might find it useful to build a collection of materials that you could send to interested parties. Chapter 53 contains many references to organizations that might be helpful in obtaining information about tests. The supplementary compact disc also contains many documents that provide varying perspectives and additional information about tests and their uses in education.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Bracken, B. A. (Ed.). (1991). *The psychoeducational assessment of preschool children.* Needham Heights, MA: Allyn & Bacon.

Donlon, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests.* New York: College Board.

Ekstrom, R. B., & Smith, D. K. (2002). *Assessing individuals with disabilities.* Washington, DC: American Psychological Association.

Haney, W. M., Madaus, G. F., & Lyons, R. (1993). *The fractured marketplace for standardized testing.* Norwell, MA: Kluwer Academic Publishers.

Helms, J. E. (1997). The triple quandary of race, culture, and social class in standardized cognitive ability testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment.* New York: Guilford Press.

✦Joint Committee on Testing Practices. (2002). *Code of fair testing practices in education.* Retrievable on AAC website at http://aac.ncat.edu.

Jones, L. V. (1971). The nature of measurement. In R. L. Thorndike (Ed.), *Educational measurement* (2d ed.). Washington, DC: American Council on Education.

Leman, N. (1999). *The big test.* New York: Farrar, Straus, & Giroux.

Messick, S. (1980). *The effectiveness of coaching for the SAT.* Princeton, NJ: Educational Testing Service.

Willingham, W. W., & Breland, H. M. (1982). *Personal qualities and college admissions.* New York: College Board.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (1990). *Predicting college grades.* Princeton, NJ: Educational Testing Service.

Zwick, R. (2002). *Fair game.* New York: Routledge Falmer.

✦ Document is included in the Anthology of Assessment Resources CD

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
X	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").